Explainable AI for Automated Multi-Disease Diagnosis: A Survey on Interpretable Deep Learning in Healthcare

1Senbagavalli M[1], 2Saswati Debnath[2], 3Shashi Kant Gupta[3]
[1,2]Post Doctoral Researcher, Lincoln University College, Malaysia
[3]Adjunct Professor, Lincoln University College, Malaysia
pdf.Senbagavalli@lincoln.edu.my, pdf.saswatidebnath@lincoln.edu.my, shashigupta@lincoln.edu.my

Abstract

Artificial intelligence (AI) has revolutionized healthcare, particularly in automating disease diagnosis through deep learning (DL). However, the opaque nature of many DL models continues to obstruct their adoption in clinical settings due to a lack of interpretability and trust. This paper presents a comprehensive survey of Explainable AI (XAI) approaches that address these concerns by integrating interpretability into diagnostic systems. We examine current literature on multi-disease diagnosis using heterogeneous healthcare data including electronic health records, lab results, and medical imaging and evaluate the effectiveness of popular XAI techniques such as SHAP, LIME, and Grad-CAM. Our survey outlines the strengths and limitations of each approach, discusses clinical relevance, and highlights directions for future research toward responsible, transparent AI in medicine.

Keywords: Explainable AI, XAI, deep learning, medical diagnosis, SHAP, LIME, Grad-CAM, healthcare AI, multi-modal data, interpretability

1. Introduction

The integration of artificial intelligence (AI), particularly deep learning (DL), into healthcare has enabled remarkable advancements in the automatic diagnosis of diseases. DL models, such as convolutional neural networks (CNNs) and transformers, are widely applied in interpreting complex healthcare data, including radiological images, structured lab tests, and unstructured clinical notes. Despite their accuracy, these models often function as 'black boxes,' providing predictions without clear reasoning, which limits their trust and adoption in real-world clinical workflows.In high-stakes environments like healthcare, interpretability is not just a desirable trait, it is essential. Clinicians must understand, trust, and validate the decisions made by AI systems. The absence of transparency in AI-based diagnostic tools hinders regulatory approval,

complicates integration into electronic health record (EHR) systems, and poses ethical concerns.

Explainable AI (XAI) has emerged as a solution to this challenge. By providing insights into model predictions, XAI enhances the credibility, accountability, and usability of AI systems in healthcare. Techniques such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and Grad-CAM (Gradient-weighted Class Activation Mapping) allow developers and clinicians to understand and validate model behavior.This survey presents a comprehensive overview of XAI methods in the context of multi-disease diagnosis using diverse data types. It identifies current gaps in research, compares existing methods, and highlights opportunities to enhance clinical applicability through explainability.

2. Survey of Explainable AI Techniques in Multi-Disease Diagnosis

The integration of **Explainable Artificial Intelligence (XAI)** into medical imaging and diagnostics has gained substantial attention in recent years. Numerous studies have explored different disease domains, data modalities, and interpretability techniques, aiming to bridge the gap between black-box models and clinical trust.

The critical innovation rests in the hybrid architecture that integrates multi-scale feature extraction from InceptionV3 and DenseNet121's dense connectivity profile, along with custom attention and feature fusion mechanisms. This development marks a breakthrough in the automated ocular disease diagnostics system as it greatly surpasses the performance of single-architecture and other hybrid models. The findings of this study highlight the technology's capability to transform the automated diagnostics landscape for ocular diseases, providing a clinically usable technology for the early detection of diabetic retinopathy, glaucoma, and age-related macular degeneration [1]. This research proposes a novel model for the deep learning-based diagnosis of nine types of gall bladder diseases. These include gallstones, abdomen and retroperitoneal pathology, cholecystitis, membranous and gangrenous cholecystitis, perforation, polyp and cholesterol crystals, adenomyomatosis, carcinoma and various gallbladder wall thickening causes. The model utilizes attention guided residual convolutional neural networks, which combines multi-scale feature extraction via dilated convolutions, attention-based feature selection, and residual connections to mitigate information loss and the vanishing gradient effect [2].

A new deep learning model has been put forward, the convolutional state space model with multi-window cross-scan (Mwinc-Mamba). Mwinc-Mamba integrates the strengths of CNN and SSM with a multi-window cross-scan approach to achieve the capture of multi-grained lesion features. This greatly enhances the capability of automated skeletal fluorosis diagnosis[3]. With full automation, the model incorporated tooth segmentation and numbering, and also provided

detailed segmentation of dental diseases. With the model, dental professionals can optimize their manual work and improve the clinical diagnosis processes. This study developed a DL approach for a thorough evaluation and diagnosis of bitewings, emphasizing the potential to improve the efficiency and precision of record filing in dental practice.[4] The detailed proposal of ABUS-Net for breast cancer diagnosis using ABUS is provided including the architecture of multi-scale feature extraction, multi-scale feature fusion, graph structure design, and GCN classification module[5].

They showed like how the proposed methods improve the accuracy of the diagnosis and the ease of interpretation through experiments with two datasets of heart sound signals. The heart sound datasets include Yaseen heart sound public dataset which is sourced from books and random internet sites, and a clinical dataset from a hospital which uses a digital stethoscope(DT) during examinations. The WCFormer method not only achieves high-accuracy diagnosis of CVDs but also possesses model interpretability enhancing the creditability of the decision-making results[6]. The architecture is based on a multi-layered analytical framework that combines local and spatial features with long-range contextual dependencies to capture complex medical image features. Furthermore, the consistency provided by the BSCRADNet mitigates the risks associated with high inter-observer variability or low visibility of the lesion. The model also showed high performance on various datasets with 94.67% accuracy on brain tumors, 89.58% on skin cancer and 90.40% on diabetic retinopathy [7].

The proposed system Suggests the application of Vision Transformers and Perceiver IO as a backbone of a hybrid AI for classification of medical images attending multi-diseases. Achieves high accuracy and low false positives across the domains of neurology, dermatology, and pulmonology. Attains up to 1.00 recall for neurology, skin, and lung conditions evaluating six diseases across benchmark datasets. Implements a real-time chatbot for diagnostic image upload and interpretation, providing automated confidence scores. Accuracy and computational efficiency ViT + Perceiver IO outperformed CNN models for the first time in these diseases[8]. This iterative process of feature extraction, clustering, pseudo-labeling, and fine-tuning continues until the PRes-SE-att-ViT model fully converges. The experimental results achieved detecting plus disease with 99.3% accuracy, 98.8% precision, 99.1% recall, and F1 score of 99%[9]. Apply multi-objective feature selection to focused the criteria of importance to a high-dimensional feature subset. Apply Harris Hawk optimization to enhance a shift from local to global search seamlessly. Employ an automated machine learning approach to streamline the iterative process of model development. Integrate tree-based pipeline optimization to merge the work of a subset of the Scikit-learn machine learning framework with a stochastic global search algorithm based on genetic programming[10].

Ref. No.	Proposed Details	Gap Identified
1	Hybrid architecture integrating InceptionV3 + DenseNet121 with custom attention and feature fusion for multi-class ocular disease diagnosis (diabetic retinopathy, glaucoma, AMD).	Needs validation across larger, diverse datasets; generalizability to more ocular diseases not explored.
2	Attention-guided residual CNN with multi-scale dilated convolutions and residual connections for gallbladder disease diagnosis (9 types).	Limited to ultrasound imaging; lacks cross-modality generalizability and real-time deployment.
3	Convolutional state space model with multi-window cross-scan (Mwinc-Mamba) integrating CNN + SSM for skeletal fluorosis diagnosis.	Needs clinical validation; scalability to other bone-related diseases unexplored.
4	Automated chart filing and tooth segmentation for bitewings using deep learning, aiding clinical dental practice.	Model limited to bitewings; lacks multimodal dental data integration and longitudinal studies.
5	ABUS-Net: GCN with multi-scale features and fusion for breast cancer diagnosis using automated breast ultrasound (ABUS).	Needs validation on larger and more diverse ABUS datasets; interpretability of graph structures not fully explored.
6	WCFormer: interpretable deep learning framework for heart sound signal analysis and cardiovascular disease diagnosis, tested on public + clinical datasets.	Requires deployment in real-time hospital systems; performance under noisy conditions not tested.
7	BSCRADNet: multi-layered framework combining local, spatial, and long-range features for medical imaging (brain tumor, skin cancer, DR).	Generalizability beyond tested datasets unclear; lacks explainability mechanisms.
8	Hybrid AI with Vision Transformers + Perceiver IO for multi-disease imaging diagnosis; chatbot integration for real-time diagnostic assistance.	Still at experimental stage; scalability and deployment feasibility in real-world clinics untested.
9	PRes-SE-att-ViT model with iterative clustering, pseudo-labeling, and fine-tuning for plus disease detection in retinopathy of prematurity.	Dataset size limited; validation in different clinical environments not conducted.
10	Multi-objective hybrid Harris Hawk optimization with AutoML pipeline for disease diagnosis and feature selection.	Limited experimental validation; lacks specific clinical application studies.

Table 1. Summary table of literature survey with gap analysis

2.1. XAI in Thoracic and Pulmonary Disease Diagnosis

A significant focus of recent research has been on the application of XAI in chest radiography (CXR). Several studies have utilized **Gradient-weighted Class Activation Mapping (Grad-CAM)** and **Local Interpretable Model-Agnostic Explanations (LIME)** to interpret predictions made by CNNs and ResNet variants on thoracic diseases such as tuberculosis and pulmonary conditions.

For instance, one model achieved a high AUC using Grad-CAM on the **ChestX-ray14** dataset, while another integrated ResNet-50 with LIME, achieving over 93% accuracy in detecting pulmonary diseases. Although these approaches enhance visual explainability, their generalizability to other modalities and real-world deployment remains limited.

2.2. Multimodal and Structured Data Integration

To improve diagnostic accuracy, recent studies have explored the fusion of **multimodal data** (e.g., images, structured EHR, pathology reports). A deep learning model integrating CNNs, RNNs, and Transformer architectures demonstrated superior performance on combined image and structured data compared to single-modal approaches. Similarly, multimodal fusion using clustering for osteoporosis detection improved interpretability. However, these models often **lack robust XAI mechanisms** or real-time validation. Other works focused exclusively on **structured data**, such as EHRs, to classify diseases like urinary tract infections using clinically relevant features. Yet, such models exclude image-based diagnostics, thus limiting the model's comprehensiveness.

2.3. XAI for Critical and Intensive Care Predictions

Explainable models have also been proposed for **ICU outcome predictions**, where time-series models like **LSTM combined with SHAP** were applied to the MIMIC-III dataset for mortality prediction. Although promising in offline evaluation, these models have yet to address **real-time deployment challenges**, such as latency, system integration, and clinician interpretability in dynamic environments.

2.4. Disease-Specific XAI Applications

Several studies have concentrated on **specific disease domains**. For example, lung cancer detection using a concept-bottleneck approach achieved an F1-score >0.9, offering more transparent explanations than SHAP or LIME. Similarly, XAI for breast cancer classification highlighted the superior visual pattern recognition of Grad-CAM on mammograms. In dermatology, **MICA** provided multi-level image-concept alignment for skin lesion diagnosis, though it lacked structured data integration. For neurodegenerative diseases, attention-based models were used in Alzheimer's diagnosis via MRI imaging, where localized brain regions were highlighted through explainability mechanisms. Despite clinical applicability, scalability across brain disorders and inclusion of additional modalities remain open challenges.

2.5. Federated and Trust-Aware XAI Models

Emerging research has considered **federated learning** and trust-aware frameworks to improve clinical relevance and privacy-preserving explainability. For example, a pediatric echocardiography model combined federated learning and XAI to ensure secure, interpretable

diagnostics. Another study validated SHAP-based explanations directly with clinical experts, enhancing trustworthiness but omitting visual or unstructured data.

2.6. Gaps and Future Directions

While the reviewed studies demonstrate progress in integrating XAI into healthcare diagnostics, several **gaps remain consistent** across the literature:

- **Modality Limitations**: Many models are restricted to specific data types (e.g., only imaging or only EHR).
- **Lack of Generalizability**: Models often focus on a single disease domain and lack validation across broader diagnostic categories.
- **Deployment Challenges**: Real-time implementation, clinician usability, and integration with hospital systems are rarely addressed.
- **Explainability Depth**: Although tools like SHAP, LIME, and Grad-CAM are used widely, most studies stop at surface-level interpretation without evaluating clinical understanding or impact.

These challenges underscore the need for **scalable, multimodal, clinically validated, and real-time explainable models**, capable of improving decision-making without compromising transparency or trust.

3. Results and Discussion

The survey reveals a growing trend toward integrating XAI into disease diagnosis systems. However, significant gaps remain:

- **Clinical Validation**: Few studies involve healthcare professionals in evaluating explanations
- **Performance Trade-offs**: Increased interpretability can sometimes reduce model accuracy
- **Standardization Issues**: Lack of common benchmarks for explainability in medical AI

Strengths of current systems include their ability to:

- Combine structured and unstructured data for holistic diagnosis
- Use multiple XAI tools to improve interpretability
- Align with ethical and legal standards such as GDPR and HIPAA

Limitations include data scarcity, generalizability, and computational overhead of XAI algorithms.

4. Conclusion

Explainable AI offers a promising path toward trustworthy, deployable deep learning systems in healthcare. As diagnostic systems evolve to handle diverse data types and complex conditions, integrating interpretability becomes crucial. This survey highlighted key techniques and case studies, showing how SHAP, LIME, and Grad-CAM are shaping transparent AI models. Future research must focus on clinical validation, user-centric design, and regulatory alignment to build systems that doctors can rely on,and patients can trust.

References

[1] Ş. Kılıç, "HybridVisionNet: An advanced hybrid deep learning framework for automated multi-class ocular disease diagnosis using fundus imaging," *Ain Shams Engineering Journal*, Jul. 2025. doi: 10.1016/j.asej.2025.103594

[2] S. Rashid, C. J. Das, A. Chauhan, G. Aggarwal, R. C. Joshi, R. Burget, and M. K. Dutta, "Self-attention-guided residual deep neural network with multi-scale dilated feature extraction for automated gallbladder disease diagnosis in ultrasound imaging," *Computer Methods and Programs in Biomedicine*, vol. 271, p. 109020, 2025. doi: 10.1016/j.cmpb.2025.109020

[3] H. Xu, Y. Wu, R. Xie, J. Xu, J. Wu, R. Wang, and Y. Tian, "Convolutional state space model with multi-window cross-scan to advance the automated diagnosis of skeletal fluorosis," *Biomedical Signal Processing and Control*, vol. 103, p. 107439, 2025. doi: 10.1016/j.bspc.2024.107439

[4] L. Cao, N. van Nistelrooij, E. T. Chaves, S. Bergé, M. S. Cenci, T. Xi, B. Loomans, and S. Vinayahalingam, "Automated chart filing on bitewings using deep learning: enhancing clinical diagnosis in a multi-center study," *Journal of Dentistry*, vol. 161, p. 105919, 2025. doi: 10.1016/j.jdent.2025.105919

[5] C. Wang, Y. Guo, H. Chen, Q. Guo, H. He, L. Chen, and Q. Zhang, "ABUS-Net: Graph convolutional network with multi-scale features for breast cancer diagnosis using automated breast ultrasound," *Expert Systems with Applications*, vol. 273, p. 126978, 2025. doi: 10.1016/j.eswa.2025.126978

[6] S. Wang, J. Hu, Y. Du, X. Yuan, Z. Xie, and P. Liang, "WCFormer: An interpretable deep learning framework for heart sound signal analysis and automated diagnosis of cardiovascular diseases," *Expert Systems with Applications*, vol. 276, p. 127238, 2025. doi: 10.1016/j.eswa.2025.127238

[7] H. C. Reis and V. Turk, "A multi-scale context-aware deep learning framework for medical disease diagnosis," *Biomedical Signal Processing and Control*, vol. 112, pt. B, p. 108558, 2026. doi: 10.1016/j.bspc.2025.108558

[8] A. Khaliq, F. Ahmad, H. U. Rehman, S. A. Alanazi, H. Haleem, K. Junaid, and E. Andrikopoulou, "Revolutionizing medical imaging: A cutting-edge AI framework with vision transformers and perceiver IO for multi-disease diagnosis," *Computational Biology and Chemistry*, vol. 119, p. 108586, 2025. doi: 10.1016/j.compbiolchem.2025.108586

[9] K. Deepthi, M. S. Josephine, and V. J. Raja, "Automated diagnosis of plus disease in retinopathy of prematurity based on transformer-based unsupervised curriculum learning," *Biomedical Signal Processing and Control*, vol. 104, p. 107521, 2025. doi: 10.1016/j.bspc.2025.107521

[10] M. Kuanr and P. Mohapatra, "A recommender system with multi-objective hybrid Harris Hawk optimization for feature selection and disease diagnosis," *Healthcare Analytics*, vol. 7, p. 100384, 2025. doi: 10.1016/j.health.2025.100384